Latent Video Diffusion Models for High-Fidelity Long Video Generation

Year: 2022 | Citations: 330 | Authors: Yin-Yin He, Tianyu Yang, Yong Zhang, Ying Shan, Qifeng Chen

---

Abstract

AI-generated content has attracted lots of attention recently, but photo-realistic video synthesis is still challenging. Although many attempts using GANs and autoregressive models have been made in this area, the visual quality and length of generated videos are far from satisfactory. Diffusion models have shown remarkable results recently but require significant computational resources. To address this, we introduce lightweight video diffusion models by leveraging a low-dimensional 3D latent space, significantly outperforming previous pixel-space video diffusion models under a limited computational budget. In addition, we propose hierarchical diffusion in the latent space such that longer videos with more than one thousand frames can be produced. To further overcome the performance degradation issue for long video generation, we propose conditional latent perturbation and unconditional guidance that effectively mitigate the accumulated errors during the extension of video length. Extensive experiments on small domain datasets of different categories suggest that our framework generates more realistic and longer videos than previous strong baselines. We additionally provide an extension to large-scale text-to-video generation to demonstrate the superiority of our work. Our code and models will be made publicly available.